**< 고객데이터 전처리 >**  
Raw데이터는 데이터 편향성도 짙고(Long tail) 원하는 고객세분화가 나올 것 같지 않아서 전처리를 하기로 함

1. RFM Score  
R 칼럼은 비교적 원하는 방향성을 갖고 있으므로 단순히 qcut을 통해서 %별로 나눔  
따라서 F, M 칼럼에 대해서 기준을 가지고 섹션을 나누기로 함  
데이터셋을 outlier를 기준으로 나눔

( MIN~이상치 이전의 데이터셋 [A] / 이상치 이후의 데이터셋~MAX [B] )

1-1. 4 Section  
  min~A.mean / ~outlier / ~B.mean / ~max  
1-2. 5 Section  
  B에서도 이상치가 있을 것이니 그것을 기준으로 또 나눠봄 [C]  
  min~A.mean() / ~outlier / ~B.mean() / ~C / ~max  
2. Robust Scaling  
   중앙값과 IQR을 사용하기 때문에 이상치가 많고, 민감한 데이터여도 이상치의 영향을 줄일 수 있다.  
3. Log Scaling  
  데이터가 급격하게 변하는 양상이므로 적합하다

**< Cluster 후보 >**  
K-mean 클러스터링을 바탕으로 진행하기로 하고

1. Loss Function이 작은 스케일링 기법을 선택하고 2. ElbowMethod에서 나온 K값을 썼을 때 3. Silhouette Score를 비교하자

|  |  |  |  |
| --- | --- | --- | --- |
|  | Loss | Elbow | Silhouette |
| 1-1 | **779** | **4** | **0.5237** |
| 1-2 | 1182 | 4 | 0.4325 |
| 2 | 1477 | 4 | 0.4621 |
| 3 | 1961 | 4 | 0.3710 |

>> 결과적으로 1-1. RFM(4) Score 로 전처리한 데이터셋 사용

**< 클러스터 확인 >**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 고객수 | R | F | M |
| 0 | 195 | **↑** | **↑** | **↑** |
| 1 | 390 | **↓** | **↓** | **↓** |
| 2 | 351 | **↓** | **↑** | **↑** |
| 3 | 532 | **↑** | **↓** | **↓** |

|  |  |  |
| --- | --- | --- |
|  | 해석 | 등급 |
| 0 | 높은 방문율과 구매율을 가졌지만 방문이 뜸해진 고객 | 재구매 유도 고객 |
| 1 | 자주 방문하지만 높은 실적을 남기진 않는 금액 | 일반 고객 |
| 2 | 최근까지도 방문하며 매장에 이익을 가져다 주는 우수 고객 | 프리미엄 고객 |
| 3 | 방문도 뜸하고 구매율이 적은 고객 | 이탈 위험 고객 |

**< 데이터 확인(평균) >**

|  |  |  |  |
| --- | --- | --- | --- |
|  | R | F | M |
| 전체 | 145 | 36 | 2964 |
| 0 | 229 | 62 | 4822 |
| 1 | 63 | 14 | 1149 |
| 2 | 57 | 80 | 7154 |
| 3 | 232 | 12 | 849 |

**< 클러스터 분포 >**

스크린샷, 라인, 도표이(가) 표시된 사진

자동 생성된 설명 스크린샷, 도표이(가) 표시된 사진

자동 생성된 설명]

< Log >

도표, 스크린샷, 다채로움, 디자인이(가) 표시된 사진

자동 생성된 설명

**< Log >**

도표, 스크린샷, 다채로움, 텍스트이(가) 표시된 사진

자동 생성된 설명

**< Box-Cox >**

**텍스트, 폰트, 스크린샷, 화이트이(가) 표시된 사진

자동 생성된 설명**